



Real-time retinal layer segmentation of OCT volumes with GPU accelerated inferencing using a compressed, low-latency neural network

SVETLANA BORKOVKINA,¹ ACNER CAMINO,² WORAWEE JANPONGSRI,¹ MARINKO V. SARUNIC,¹ AND YIFAN JIAN^{2,3,*}

¹Department of Engineering Science, Simon Fraser University, Burnaby, Canada

²Casey Eye Institute, Oregon Health & Science University, Portland, OR 97239, USA

³Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239, USA

*jian@ohsu.edu

Abstract: Segmentation of retinal layers in optical coherence tomography (OCT) is an essential step in OCT image analysis for screening, diagnosis, and assessment of retinal disease progression. Real-time segmentation together with high-speed OCT volume acquisition allows rendering of *en face* OCT of arbitrary retinal layers, which can be used to increase the yield rate of high-quality scans, provide real-time feedback during image-guided surgeries, and compensate aberrations in adaptive optics (AO) OCT without using wavefront sensors. We demonstrate here unprecedented real-time OCT segmentation of eight retinal layer boundaries achieved by 3 levels of optimization: 1) a modified, low complexity, neural network structure, 2) an innovative scheme of neural network compression with TensorRT, and 3) specialized GPU hardware to accelerate computation. Inferencing with the compressed network U-NetRT took 3.5 ms, improving by 21 times the speed of conventional U-Net inference without reducing the accuracy. The latency of the entire pipeline from data acquisition to inferencing was only 41 ms, enabled by parallelized batch processing. The system and method allow real-time updating of *en face* OCT and OCTA visualizations of arbitrary retinal layers and plexuses in continuous mode scanning. To the best of our knowledge, our work is the first demonstration of an ophthalmic imager with embedded artificial intelligence (AI) providing real-time feedback.

© 2020 Optical Society of America under the terms of the [OSA Open Access Publishing Agreement](#)

1. Introduction

Ophthalmology is at the forefront of the AI revolution in medicine. Image-based retinal disease screening with deep learning has been a hot topic in recent ophthalmology research. During the past 5 years, the group at Google's DeepMind made significant contributions to the diagnosis and referral of diabetic retinopathy (DR) using fundus photography [1] and optical coherence tomography (OCT) [2], the two most common imaging modalities found in ophthalmic clinics worldwide. Following these groundbreaking works, significant research efforts have been dedicated to AI-assisted diagnosis solutions for several vision-threatening retinal diseases such as age-related macular degeneration (AMD) [3–6], DR [7,8], glaucoma [9], and retinopathy of prematurity [10], among others. In addition to becoming a potential screening tool, AI-assisted segmentation technologies can help in the quantification of biomarkers associated with disease progression [11–17]. However, all of the above-mentioned research only trained and applied AI methods in post processing after the images had been acquired. There is an opportunity to apply deep learning during the image acquisition to assist the operator's judgment of image quality prior to acquisition, identify potential areas of disease and provide real time feedback. These capabilities could significantly improve the imaging success rate and potentially diagnosis accuracy.

OCT is a non-invasive, three-dimensional medical imaging technology that has become the standard of care for the evaluation of macular disease. OCT is increasingly being used not just for cross-sectional visualization of tissue but also for depth resolved *en face* imaging of arbitrary retinal depths [18–20]. OCT-angiography (OCTA) particularly exploits the benefits of *en face* retinal visualization to assess the integrity of retinal capillary networks. OCT-OCTA is now also used to detect subclinical disease [21,22] and obtain quantitative metrics of disease severity in structure and blood flow simultaneously, encompassing layer thickness [23], ellipsoid zone defects [24], drusen area [25], vessel density [26], among others. The accurate *en face* visualization of separate retinal layers with OCT and retinal plexuses with OCTA relies on the correct segmentation of the cross-sectional B-scans composing OCT volumes.

Because accurate layer segmentation is essential for appropriate clinical interpretation, this has been an active area in OCT research. Many conventional image processing methods have been proposed in the literature to segment retinal layer boundaries, including active contour [27], support vector machine [28], and graph-based methods [29]. Machine learning, and specifically deep neural networks, have also been used to perform retinal segmentation in OCT images. Roy *et al.*, [14] used ReLayNet, a fully-convolutional neural network, to segment retinal layers and fluid in macular OCT images. Fang *et al.*, [30] used a combination of deep learning and graph search to segment nine retinal layer boundaries in AMD patients. Most of these retinal layer segmentation methods are too slow to segment retinal layers in real-time, either due to algorithmic or implementation complexities.

With increasing OCT acquisition speed, real-time volumetric OCT has become a reality and real-time segmentation of retinal layers has become more desirable. In addition to speeding up the OCT volume data processing workflow and providing prompt diagnosis feedback, real-time OCT data processing and visualization of *en face* retinal layers is also a key requirement of successful imaging in the clinical setting. For example, real-time retinal layer segmentation can provide effective focus control and the direct feedback of aberration correction performance with image-guided adaptive-optics (AO) techniques. It could also be used in ultra-wide field of view OCT systems to dynamically control the reference arm length to match with the retinal curvature. For surgical applications, real-time segmentation of OCT features with deep learning could assist intraocular surgery and antiangiogenic therapy in a similar way deep-learning based real-time segmentation has aided prostate biopsy [31] and fetal standard scanning [32] with ultrasound, as well as cardiac [33] and fetal brain [34] volume extraction with MRI.

The first barrier for the real time retinal layer segmentation is the Fourier-domain (FD) OCT image processing itself, which is an inherently computationally intensive process. Transformation of interferometric fringes into B-scan images from high-speed FD-OCT systems far exceeds the computation capability of the state-of-the-art central processing unit (CPU). In order to keep up with the ultra-high speed OCT acquisition speed, we developed a parallel computing OCT image processing platform enabled by the graphics processing unit (GPU) that allows real-time OCT processing [35]. Our custom GPU pipeline can perform OCT processing at 2.24 MHz axial scan rate and has accomplished streaming of flow contrast *en face* images extracted from the selected region on speckle variance OCTA (svOCTA) in real-time [36].

With real-time OCT processing capability, some implementations have demonstrated real-time axial tracking based on OCT cross sectional images [37–39]. The axial position of the retina in an OCT B-scan was determined by the brightness of layers, a method that allows real-time focus optimization and visualization of OCT *en face* views. However, real-time axial tracking does not provide retinal layer specific information that is based on anatomical structures. Recently, Janpongsri *et al.*, [40] demonstrated pseudo-real-time retinal layer segmentation for high-resolution sensorless adaptive-optics optical coherence tomography angiography (SAO-OCTA) with a segmentation method based on Dijkstra's algorithm. However, the segmentation time of this method strongly scales with the number of the retinal layers to be segmented and the

size of the image. For instance, segmentation of 7 retinal layers requires more than 20 ms which is insufficient for high speed OCT that usually acquires B-scans at rates of a few hundreds of frames per second. Moreover, retinal layers can change drastically their thickness and delineation for different retinal landmarks and diseases. Artificial intelligence has more potential than classical image processing methods to infer successfully the diversity features described by retinal boundaries.

In this work, a modified U-Net architecture was trained for segmentation of eight retinal layer boundaries with data annotated by an offline and manually corrected graph-cut segmentation. For real-time inferencing, we generated a low complexity network (U-NetRT) by compressing the trained network's graph with specific layer pruning and fusions that reduced the computational load and maintained the original network's performance. We also used dedicated hardware (Tensor Cores) operating in parallel and designed to further accelerate the massive multiplication of matrices involved in deep learning, which maximized the performance of the GPU. We demonstrate high accuracy of layer boundary segmentation and *en face* visualization of arbitrary layer/slab simultaneously with OCT image acquisition.

2. Materials and methods

The real-time deep learning segmentation workflow is summarized in Fig. 1. First, ground truth layer segmentation of the B-scans in the training set were generated by an established graph-search method, which could be manually corrected in the event of segmentation errors. The available data was augmented by applying random contrast and intensity to each input B-scan, and a neural network was trained to obtain the probability of pixels belonging to retinal layers, the vitreous humor and the choroid. The originally-trained network architecture was then further compressed by TensorRT and the precision of weights was reduced by a scaling factor determined in a calibration process, reducing the computational burden in order to perform real-time inferencing. Real-time rendering of *en face* projections of arbitrary layers simultaneously (see ganglion cell layer and outer plexiform layers in Fig. 1) was achieved. These steps will be explained in detail in the following sections.

A total of ten volumes were used in this work from two participants. To improve signal-to-noise ratio and reduce speckle noise, a rolling average was applied to all groups of 10 adjacent B-scans. Averaged B-scans with blurred layers due to eye motion were excluded from the dataset, resulting in the total size of the dataset to be 5234.

The study participants were recruited and imaged at the Casey Eye Institute at the Oregon Health & Science University (OHSU). The imaging protocol was approved by the Institutional Review Board/Ethics Committee. The research adhered to the tenets of the Declaration of Helsinki.

2.1. System description

High-resolution OCT/OCTA volumetric images were acquired from the macula of healthy subjects with a prototype adaptive-optics spectral-domain OCT system built at the Casey Eye Institute [41]. The system operated at 250 kHz A-line scan rate (600 Hz B-scan rate), with an axial resolution of 2.6 μm in tissue (assuming a 1.33 index of refraction) and an optical lateral resolution of 5.6 μm . Cross-sectional images of size 400×1024 pixels were acquired at 400 adjacent lateral positions sampled over a field of view of 3×3 mm. The computer workstation used in this work consisted of an Intel i9 9900k CPU, an NVIDIA GeForce RTX 2080 Ti GPU, and 32 GB of RAM.

2.2. Network architecture

We selected the popular U-Net architecture developed by Ronneberger *et al.*, [42], which has proven to produce highly accurate segmentations when training data is limited. U-Net is an

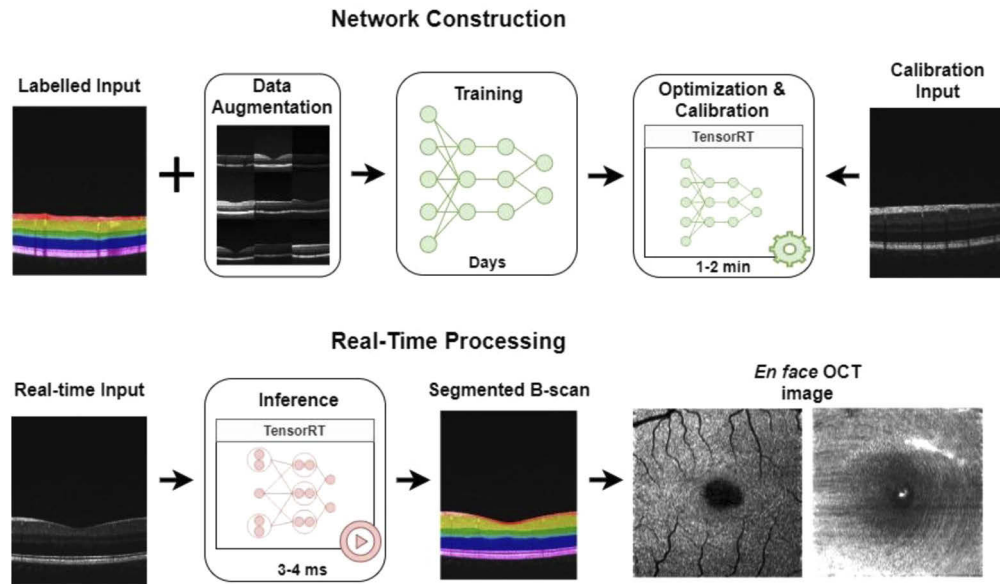


Fig. 1. Summary of deep learning segmentation workflow. After generating the ground truth and pre-processing the cross-sectional B-scans (including augmentation of B-scans available), a neural network was trained to segment seven retinal layers. An optimization process cropped the network size and a calibration that took into account the dynamic range of activations was used to reduce the precision of network weights in order to perform inference of averaged groups of 10 OCT B-scans simultaneously with acquisition. After layer boundaries had been defined, *en face* projections of the volumetric data could be produced.

end-to-end fully convolutional neural network with two symmetric paths: the contracting path to capture the context of an image, and the expansive path to precisely localize image features. In the original implementation of U-Net, each stage between pooling operations consisted of two consecutive convolutional layers. Here, aiming to reduce network size and thus inferencing time, we used one convolutional layer per stage (Fig. 2) for our modified U-Net architecture. The contracting path consisted of four repeating blocks, each containing a 3×3 convolutional layer followed by a batch normalization layer, a rectified linear unit activation function (ReLU), and a 2×2 max pooling layer with stride 2. After the first convolution, the number of feature channels is 64. This number then doubles with each consecutive convolution. The four blocks in the expansive path perform upsampling using 2×2 deconvolution layers, then concatenate the result with the corresponding feature map from the contracting path, and perform a 3×3 convolution followed by batch normalization and ReLU. The final layer in the network is a 3×3 convolution layer with ReLU activation function, which outputs the probability of a pixel to belong to each of the 8 output classes.

2.3. Training

Ground truth was generated separately for each B-scan by an automated graph cut algorithm, manually corrected in ITK-Snap [43]. We modified the shortest-path graph cut retinal layer segmentation method based on Chiu *et al.*, [44] and its open source implementation in the Caserel software. The method used Dijkstra's algorithm by iteratively searching the lowest weight path connecting the two end nodes of a layer boundary. The modified Dijkstra's algorithm was implemented with a CPU routine in Boost C++ Libraries.

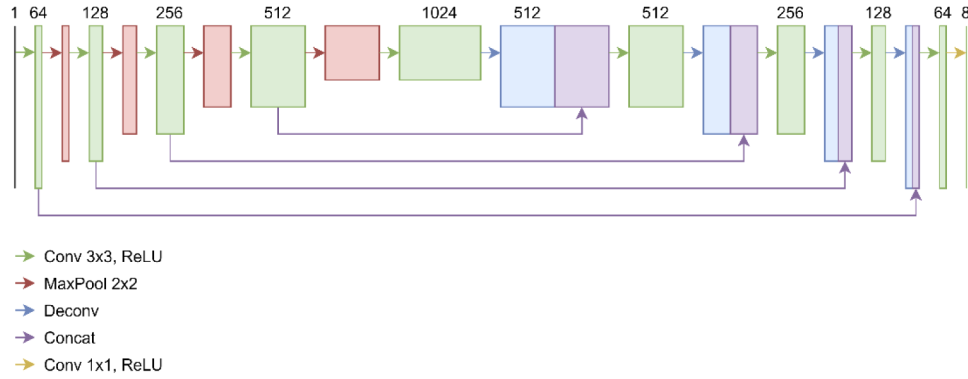


Fig. 2. Representation of network architecture of the modified U-Net trained to segment retinal layers. The number of feature channels is specified for each stage. This network was further compressed, as explained in Section 2.5, in order to be used for real-time inference simultaneously with data acquisition.

In both training and inferencing stages, we performed two pre-processing steps to homogenize the training and test data (Fig. 3). Firstly, we took the running average of all groups of 10 adjacent B-scans that were processed and presented in linear scale to improve the signal-to-noise ratio. B-scans with layers significantly blurred after averaging due to eye motion were excluded from the training and test datasets. Secondly, we increased the contrast of the averaged image by saturating the top and bottom 1% of pixel intensities and performed contrast stretching for intensities in between. Both steps were implemented using CUDA to increase throughput by taking advantage of parallelization. To ensure that no B-scans in the training and testing datasets belonged to the same volumetric acquisition, two out of ten available volumes were selected for testing, and eight were used for training. After eliminating poor-quality B-scans due to eye motion, the resulting training and testing datasets contained 4072 and 1162 images, respectively. The sizes of training and testing datasets were chosen empirically. The network was trained for 20 epochs using Adam optimizer with a learning rate set to 10^{-5} . To reduce dependencies on brightness and contrast of a B-scan, each training image at every epoch was augmented with a random contrast/brightness coefficient.

After segmentation, each pixel was assigned to one of 8 possible labels corresponding to vitreous, inner limiting membrane and nerve fiber layer (ILM/NFL), ganglion cell layer and inner plexiform layer (GCL/IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), photoreceptors and retinal pigment epithelium (PR/RPE), and the choroid.

2.4. Loss function

The network was trained using a composite loss function consisting of weighted logistic loss and dice loss as Roy *et al.*, have demonstrated that using the join dice and weighted logistics loss results in the superior performance compared to using only dice or logistics loss [14]. The overall loss function was defined as Eq. (1):

$$L_{\text{overall}} = 0.5 * L_{\text{dice}} + L_{\text{log}} \quad (1)$$

where L_{dice} is the dice loss (Eq. (2)) and L_{log} is the weighted logistic loss.

$$L_{\text{dice}} = 1 - \frac{2 \sum_{x \in \Omega} p_l(x) g_l(x)}{\sum_{x \in \Omega} p_l^2(x) + \sum_{x \in \Omega} g_l(x)} \quad (2)$$

The weighted loss function was used in order to make the network more sensitive to the layer boundaries and compensate for imbalanced classes. The weighted logistic loss is defined in

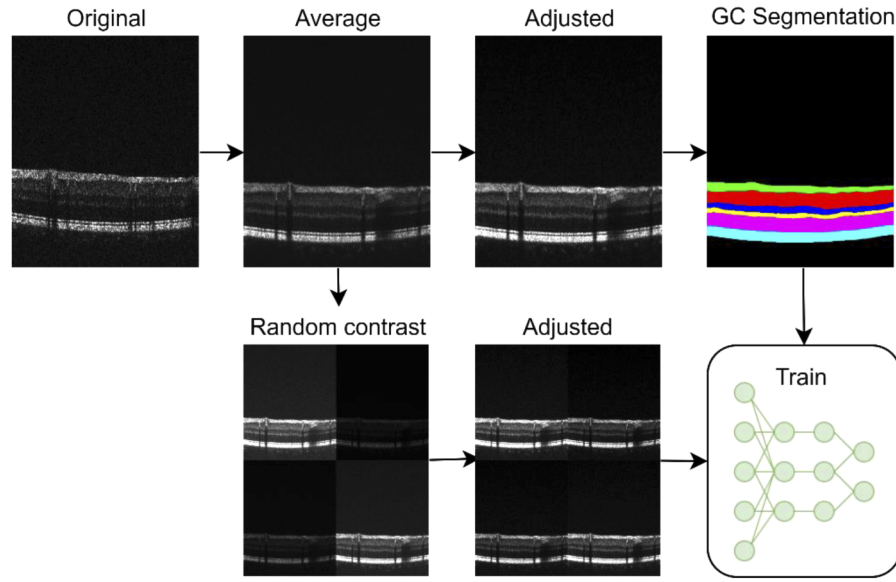


Fig. 3. Pre-processing framework of B-scans prior to training of the neural network. Ten adjacent B-scans are first averaged to increase the signal-to-noise ratio. Data was augmented by introducing random brightness adjustments in order to account for all different signal intensities. Contrast was then adjusted for original and augmented data saturating the bottom and top 1% of pixel intensities. Then, the data and ground truth obtained by graph-cut (GC) segmentation were fed to a U-Net type neural network for training.

Eq. (3):

$$L_{log} = -\sum_{x \in \Omega} w(x) g_l(x) \log(p_l(x)) \quad (3)$$

where $w(x)$ is the weight associated with a particular pixel, $g_l(x)$ is a vector containing 1 for the correct label and 0 for others, and $p_l(x)$ is the estimated probability for each label of this pixel. To encourage the network to learn layer boundaries, the regions of transition between layers are set to have a greater contribution. In addition, background regions above and below the retina weigh less than the pixels belonging to the retinal layer classes. The final weighting scheme is implemented in Eq. (4):

$$\omega(x) = 1 + 10 * I(\nabla l(x) \vee 0) + 5 * I(l(x) = L) \quad (4)$$

where I is an indicator function and ∇ is the gradient operator.

2.5. Real-time inference

Once a neural network is trained, it is possible to draw inference from the network in a matter of milliseconds using GPU acceleration. To achieve significant reduction in inferencing time, we took advantage of NVIDIA's TensorRT, a platform for high-performance deep learning inference on NVIDIA GPUs. TensorRT allows acceleration of networks trained in all popular frameworks, such as TensorFlow, Caffe, or MATLAB. It performs various optimizations on a network to improve speed performance during inference by optimizing the network graph, reducing inference precision mode, and selecting kernels specific to the GPU to maximize performance.

The purpose of graph optimization is to make the network smaller while preserving its performance. Our implementation of a modified U-Net contained 163 layers. A first optimization step pruned the layers that had no impact on the performance, which reduced the network size

down to 59 layers. Then, layer fusion was performed, further simplifying the network. The layer fusion step detected layer patterns suitable for optimization in the graph and then fused them into one. For our U-Net implementation, the following layer fusions took place:

1. Convolution and scale layers (bias addition and batch normalization) by adjusting convolution weights to result in the same computation.
2. Convolution and ReLU activation by computing ReLU function in the same kernel as the convolution.
3. Scale and activation (Bias addition and ReLU).

After the layer fusion step, 32 layers were removed, resulting in a graph with 27 layers. The final graph optimization step consisted of removing concatenation layers, which further reduced the final number of layers to 23. TensorRT is capable of performing other types of graph optimizations, but we have only discussed here the ones applied to our network. Owing to TensorRT graph optimizations, the network size was reduced from 163 to 23 layers.

In addition, neural network parameters were represented as 32-bit floating point (FP32) during training, but TensorRT supports different precision modes: FP32, FP16, and INT8 (8-bit integer). Choosing a lower precision mode can dramatically reduce inferencing time, so we used the lowest precision mode available, 8-bit integer, to achieve the fastest possible inference. The INT8 type has a much lower dynamic range compared to FP32, so to avoid accuracy loss from precision reduction, FP32 needs to be correctly mapped to INT8 with the use of calibration data to determine the dynamic range of each tensor in the network. Calibration data contains samples of representative input to the network, and need to come from the same distribution as the test/inference data. In this work, 10 B-scans were selected from the training data to cover various sections of the retina and scanning positions.

After TensorRT completed all the above-mentioned optimizations, a runtime engine allowed us to perform high-speed inference.

2.6. Tensor cores

Deep learning is a very computationally intensive task, which makes selecting suitable hardware essential for high-performance projects. Training the same network on a multi-GPU machine could reduce training time from days to hours, allowing developers to iterate through network architectures towards the best implementation relatively quickly. With the recent growth of artificial intelligence, a number of technologies have been introduced to accelerate all stages of deep learning. One of the recent innovations greatly improving machine-learning performance are NVIDIA's Tensor Cores. Tensor Cores perform multiply-accumulate calculations on a 4×4 matrix in a single operation. As convolutional neural networks commonly contain large numbers of multiply-accumulate instructions, accelerating this operation leads to significant improvement in network performance. The theoretical speedup is 8 times, with an average of 4 for a typical user. In the project, we used a NVIDIA 2080 Ti GPU that has 554 Tensor cores to further accelerate inferencing of the retinal layer segmentation.

2.7. Post-processing

The output of the neural network is a semantically segmented image, with each pixel belonging to one of the 8 possible classes. Layer boundaries were then extracted from the segmentation map and used to generate the *en face* projection between the desired layers. As neural network semantic segmentation is anticipated to have an associated amount of error, a small number of pixels is expected to be mislabeled. The A-lines containing these errors could be detected automatically because the disposition of the labels assigned along the depth direction did not follow the proper order of retinal anatomy. This is particularly common close to the fovea, as

the layers between ILM and OPL come together in this region. If a valid layer boundary cannot be detected in a section of a layer, we applied linear interpolation to fill the gap to provide a continuous boundary. Figure 4 demonstrates the steps of detecting and correcting missing layer boundaries, showing segmentation with and without the correction of retinal layers.

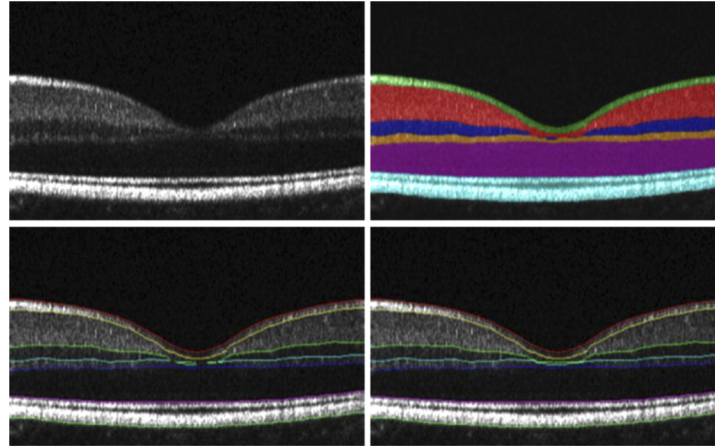


Fig. 4. Detection and correction steps of segmentation inaccuracies. (a) Original B-scan. (b) Segmentation with missing layer boundary for INL (blue). (c) Detected layer boundaries with a gap. (d) Layer boundaries after linear interpolation.

3. Results

3.1. Network performance

We evaluated the performance for the original U-Net and the modified U-Net, which only has half of the original convolutional layers. Both networks showed good agreement with ground truth (Fig. 5) and their performances were evaluated using the Dice coefficient for each B-scan (Table 1). The accuracy was comparable to the results reported by other teams using neural networks for retinal layer segmentation [14]. Reducing the number of convolutional layers of the original U-Net led to a relatively modest deterioration in Dice coefficient for our dataset, which was 0.005 in the worst-case scenario. After TensorRT optimization and reduction of precision to INT8, accuracy loss was negligible for U-NetRT compared to the modified U-Net with the largest Dice coefficient difference (0.0023) occurring at OPL layer. Dice coefficients of U-NetRT were identical to the modified U-Net prior to TensorRT for FP32 and FP16 precisions.

Table 1. Dice coefficients representing the performance of the original U-Net and the modified U-Net on the test set for the eight layer labels produced. ILM – Inner limiting membrane. NFL – Nerve fiber layer. GCL – Ganglion cell layer. IPL – Inner plexiform layer. INL – Inner nuclear layer. OPL – Outer plexiform layer. PR – Photoreceptors. RPE – Retinal pigment epithelium.

	Vitreous	ILM & NFL	GCL & IPL	INL	OPL	ONL	PR & RPE	Choroid
U-Net	0.997	0.963	0.980	0.965	0.948	0.985	0.985	0.989
Modified U-Net	0.996	0.960	0.977	0.961	0.943	0.984	0.983	0.988

3.1.1. Brightness/contrast adjustment effect

During the acquisition process, adjusting brightness and contrast of the B-scans is often desirable to improve visibility of certain features. Significantly increasing brightness results in saturation

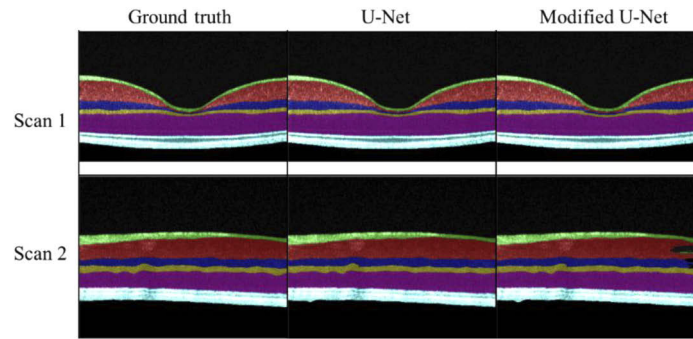


Fig. 5. Representative scans visualizing the agreement of the U-NetRT inference with the classification made by the originally trained U-Net architecture, as well as the ground truth.

of the brightest pixels, resulting in an increased number of labelling errors. In order to improve the robustness of the network when brightness and contrast are significantly different from the settings used during the acquisition of training data, we added a random augmentation step to the training process. Table 2 compares brightness and dice coefficient for two networks: the first trained on B-scans with only contrast adjustment applied, and the second trained on B-scans with a random brightness/contrast coefficient applied followed by contrast adjustment. It can be seen that the network trained on randomly augmented images performed significantly better when B-scans were augmented with a high brightness/contrast coefficient.

Table 2. Comparison of dice coefficients between two networks trained on different data when tested on images with a high brightness/contrast coefficient. Network 1 was trained on images with default brightness, while the Network 2 was trained on images with random brightness and contrast applied.

	Vitreous	ILM & NFL	GCL & IPL	INL	OPL	ONL	PR & RPE	Choroid
Network 1	0.983	0.903	0.943	0.933	0.900	0.973	0.979	0.830
Network 2	0.995	0.957	0.977	0.961	0.943	0.984	0.982	0.988

3.2. U-NetRT real-time performance

As inference time can vary for the same network and input size due to the non-deterministic nature of GPUs and non-real-time operating system, we tested the inferencing speed by running 1162 test B-scans for 100 iterations using Tensorflow on a GPU and TensorRT for each of the three available precision modes and two network architectures (Table 3). The inferencing time of the modified U-Net was reduced by 18 times after introducing TensorRT and reducing precision of network parameters (Table 3).

Table 3. Comparison of inferencing time (mean \pm standard deviation) for the two network architectures (modified U-Net and U-NetRT after architecture optimization) using TensorFlow and TensorRT with network parameters of different precision modes (floating point 32 and 16, and integer 8).

	TensorFlow	TensorRT, FP32	TensorRT, FP16	TensorRT, INT8
Original U-Net	75.66 \pm 4.87 ms	23.52 \pm 1.27 ms	6.33 \pm 0.28 ms	4.65 \pm 0.56 ms
Modified U-Net	62.46 \pm 2.87 ms	13.64 \pm 0.40 ms	4.16 \pm 0.41 ms	3.51 \pm 0.41 ms

We combined the retinal layer segmentation, pre- and post-processing steps into our real-time GPU accelerated OCT acquisition and processing software OCTViewer. In order to maintain

high speed throughout and minimize latency, the acquisition of raw OCT B-scans, transfer from host to GPU memory, and processing was synchronized with a parallelized batch scheme as shown in Fig. 6(a). Batches contained 10 cross-sectional B-scans, each formed by 400 A-lines acquired in the fast scanning direction and corresponding to 10 adjacent positions in the slow scanning direction. Batch sizes of only 10 frames allowed high throughput data transfer and processing, relatively low latency, and ensured good signal to noise ratio (SNR) when averaged for segmentation. The acquisition of batches containing 10 frames took 16.6 ms, data transfer from frame grabber 2.8 ms, OCT data processing 3.5 ms, pre-processing 0.2 ms, inference 3.5 ms and post-processing 0.11 ms (110.1 μ s for layer detection and 2.4 μ s for linear interpolation correction), for an overall latency of 41 ms (Fig. 6). This latency was considerably smaller than reported in previous literature [45].

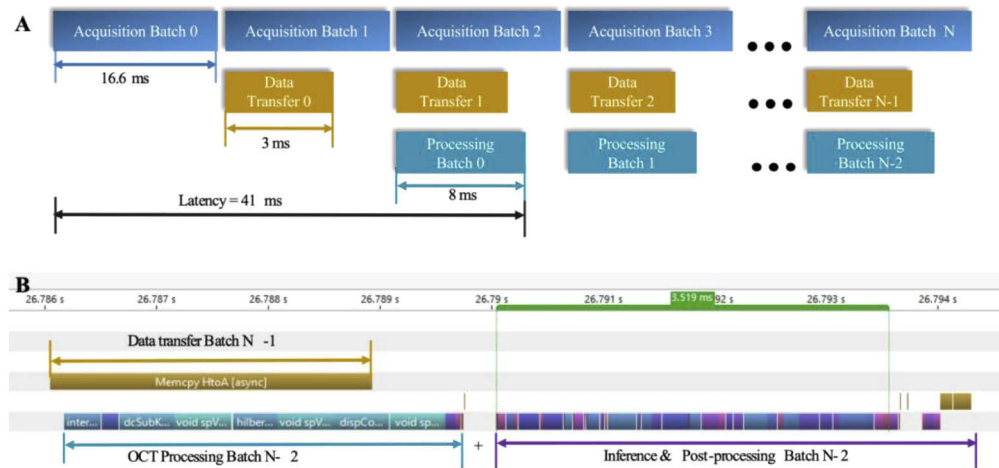


Fig. 6. Profiler for B-scan segmentation. (a) After data was acquired in batches of 10 frames (blue segments), it was transferred from host to GPU memory in synchronization with the acquisition of the next batch (orange segments). Then, OCT signal processing, image pre-processing, inference, and post-processing was performed during the acquisition of the following batch (cyan segments). (b) Processing encompasses OCT data processing, OCT image pre-processing, inference with U-NetRT and post-processing, yielding an approximate total latency of 41 ms.

Our U-NetRT exhibited a high segmentation accuracy for all layer boundaries at different eccentricities within the parafovea, including the foveal dip (Fig. 7). The pixel classification showed excellent agreement with the anatomical distribution of layers, not showing ‘islands’ of misclassified pixels, yielding layer boundaries that were continuous and smooth.

Figure 8 is a screen capture of our OCTViewer software during an *in vivo* imaging session with real-time retinal layer segmentation enabled (Visualization 1). Layer boundaries and the OCT projections of arbitrary retinal depths could be visualized simultaneously with a pupil view used to guide alignment. Due to screen size limitations, although all 6 retinal layers were segmented in real time, OCTViewer only displays four *en face* projections at a time; layer boundaries for those layers are shown on the B-scan. Real-time segmentation allowed the operator to exert better judgement of the quality of images prior to saving the OCT data. It also carries potential to represent in real-time any depth-specific retinal pathology. The segmentation was correct for various SNR and fields of view (FOV) found during *in vivo* imaging sessions (see Visualization 2 and Visualization 3). Errors were observed during head and eye movements causing pupil misalignment and partial vignetting, but were accurate again after the eye re-fixated or the alignment was corrected by the operator after the blink (Visualization 1).

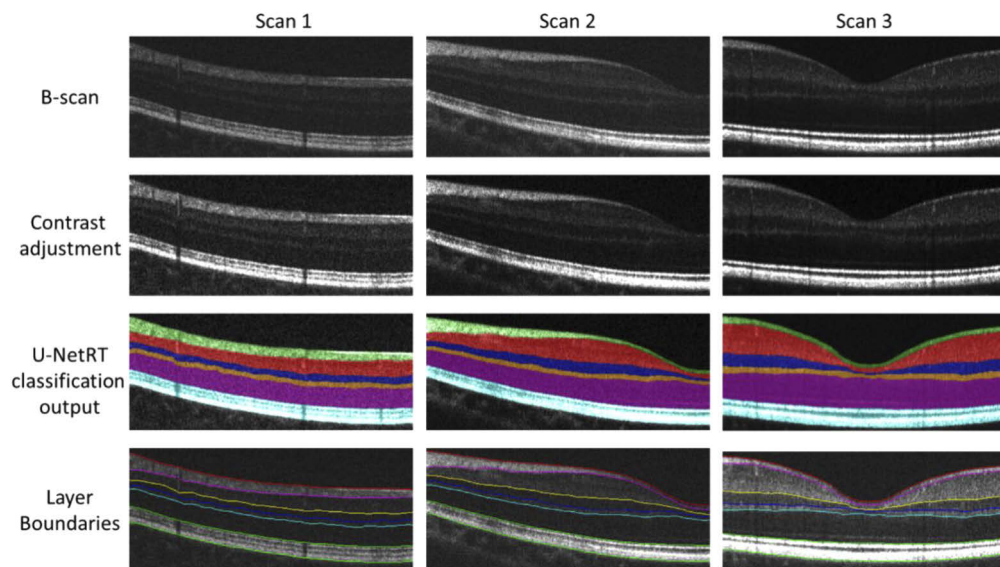


Fig. 7. Representative results of the real-time deep-learning inference processing using our U-NetRT architecture. Four steps are shown in order of occurrence: acquisition of a B-scan, contrast adjustment, layer classification of pixels, and extraction of the layer boundaries.

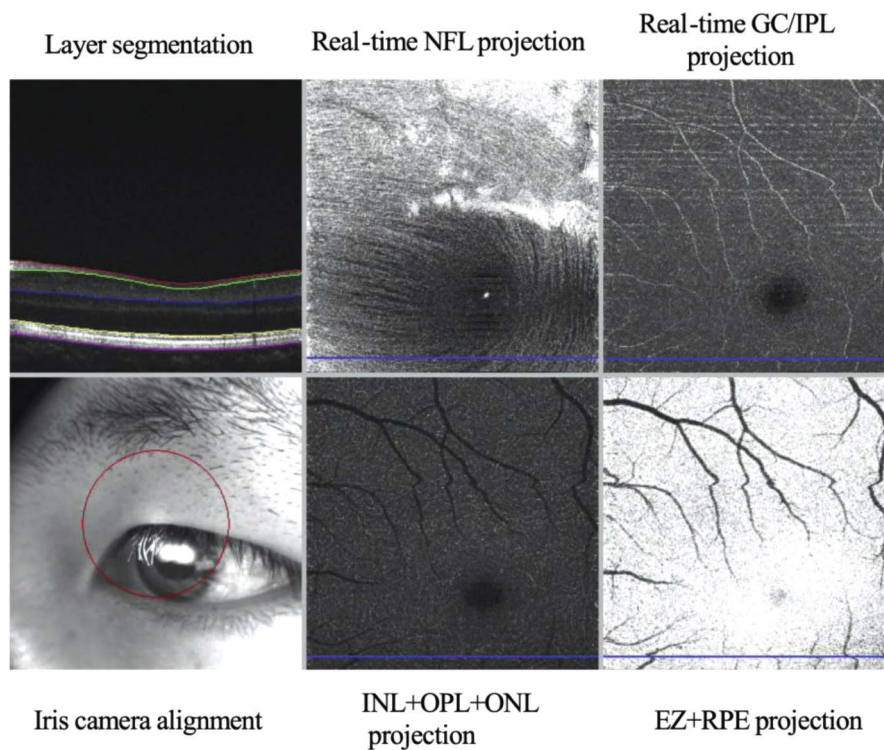


Fig. 8. Screen capture of the OCTViewer software representing *en face* visualizations of arbitrary retinal layers simultaneously with data acquisition (see [Visualization 1](#)), with a total latency of 41 ms lapsed after acquisition of the batch of frames by the spectrometer. The horizontal blue line represents the position of the B-scan visualized on the upper-left panel.

4. Discussion and conclusions

We have demonstrated a real-time retinal layer segmentation method for OCT using state of the art GPU hardware features and TensorRT to compress the neural network and reduce the computational burden while maintaining the same performance. The original neural network's architecture was compressed from 163 to 23 layers in order to reduce U-Net's inferencing time while preserving the performance above 94% for all layers. The speed of U-NetRT inferencing (3.5 ms) was reduced by 21 times with respect to the original U-Net (75.6 ms), not accounting for 0.2 ms and 0.11 ms of pre- and post-processing. The pre-processing averaging of ten adjacent B-scans was performed anticipating inaccuracies due to SNR of single B-scans, but did not incur in significant additional latency. The tests done in real time demonstrated robustness after instances of blinking and microsaccadic motion ([Visualization 1](#)), as well as in various sizes of FOV and retinal eccentricities ([Visualization 2](#) and [Visualization 3](#)).

While the training stage of neural networks typically requires large datasets and time-consuming manual segmentation of ground truth, this network was trained with only eight volumetric acquisitions from healthy eyes, each containing 400 B-scan positions within the parafovea. Data augmentation techniques allowed producing new B-scans to improve the network's accuracy without requiring new labeling; rolling averaged was applied for more efficient utilization of the available training dataset. The use of a graph cut algorithm in the generation of ground-truth itself was also a factor that facilitated training, since the operator only needed to intervene for manual corrections of the deficient frames, excluding the ones affected by eye motion. Still, the network cannot be applied yet on acquisitions encompassing the optic disc, significantly larger fields of view, or pathological characteristics such as macular edema, drusen, or layer dystrophy. We tested a retinal landmark and FOV approaching the optic disc, not used for training ([Fig. 9](#)). Here, segmentation was accurate until the boundaries approached the peripapillary area, where drastic changes in the boundary gradient owing to reduced contrast and large vessel shadows caused segmentation errors. Nevertheless, based on the previous literature we are confident that with additional training the neural network should be able to generalize successfully new patterns characteristic of different retinal landmarks and diseases.

There are some limitations of the method. Because we average 10 adjacent B-scans to improve the image quality and ensure the success of segmentations, single frame segmentations were not successful ([Fig. 10](#)) due to insufficient SNR. The method also fails when large ocular motion or blinks occur in between the average B-scans ([Fig. 10](#)). In addition, overall signal attenuation due to defocus or anterior segment opacities (e.g. cataracts) or defocus could yield an SNR too low to guarantee successful segmentation. In spite of this, the algorithm was successful across macular vessel shadows, suggesting it should also be robust to localized shadows from vitreous floaters of similar caliber. We also showed that errors in segmentation appear during microsaccadic transients in which the eye drifts away from the fixation target, which are rapidly fixed once fixation is recovered.

It can be noticed in [Fig. 6](#) that the inference process does not occur until after acquisition of the two following batches, incurring in additional latency due to the parallelized intermediate data transfer step. This configuration, as mentioned before, was preferred in order to allow the software to increase overall processing speed by effectively hiding the data transfer time at the expense of slightly longer latency. With the 250 kHz A-scan rate, we could reduce the latency by 16 ms if we either serialized the data transfer and processing or reduced the batch size by half, however we chose the batch size of 10 B-scan to avoid any potential frame skipping and sufficient frame averaging for segmentation. It is also possible that in the near future, with the PCIe 4.0 standard, to transfer data between host and GPU at a much faster speed, which would allow serialized data transfer and processing. Furthermore, a device-to-device (D2D) data transfer is also an available option to reduce the overhead. For example, with the AlazerTech digitizers or Bitflow frame grabbers it is feasible to transfer the acquired interferogram from the

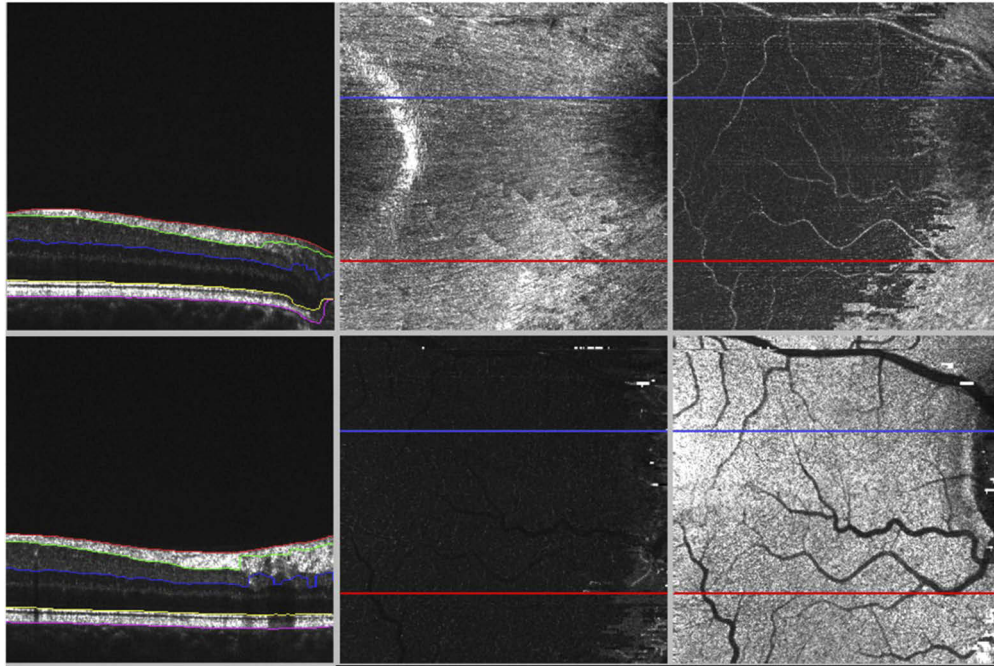


Fig. 9. Screen capture of the OCTViewer software representing segmentation performance in the area between optic disc and macula. Four *en face* OCT projections of the RNFL, GCIPL, INL-OPL-ONL and EZ-RPE slabs are visualized. The horizontal blue line represents the B-scan in the upper left corner and the red line represents the one in the lower left corner. Although this field of view was not used in training, segmentation was correct except in the area inside the optic disc.

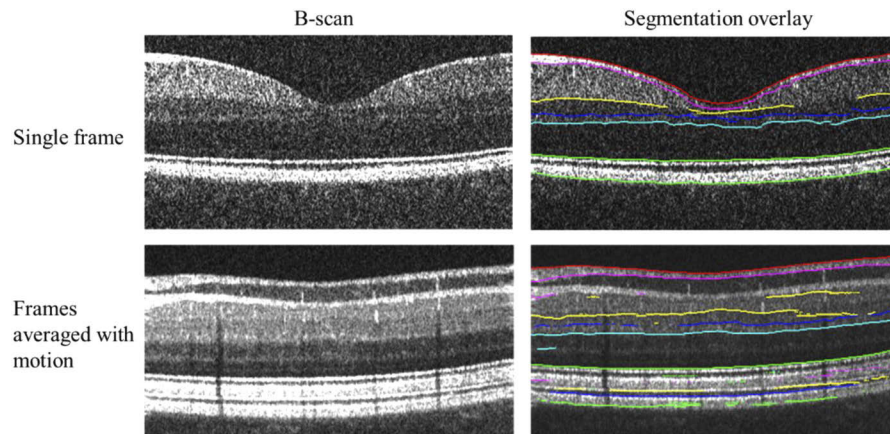


Fig. 10. Representative examples of neural network segmentation errors. Single-frame B-scans (upper row) without frame averaging did not guarantee the signal-to-noise ratio required to delineate relatively dark boundaries such as the inner nuclear layer with the inner and outer plexiform layers, yielding ‘broken’ layer boundary graphs. Although averaging of ten B-scans over five adjacent B-scan positions increased the signal-to-noise ratio, the frames found at the proximity of positions of large bulk motion artifacts presented distinct layers overlapping at the same axial depths, precluding with the performance of the neural network.

digitizer/frame grabber card directly to GPU without copying to host RAM first. However this feature requires NVIDIA Quadro or Tesla GPUs which can be much more expensive than the consumer ones used in this manuscript.

Dramatic acceleration of neural network-based segmentation is possible with little investment from researchers and developers by optimizing the neural network using existing software solutions (TensorRT) and high-end consumer GPU (GeForce RTX 2080 Ti). We were able to achieve average inference time of around 3.5 ms per B-scan, which allowed for real-time segmentation of the retinal layers during an imaging session. In order to further reduce the total latency after frame batch acquisition, two separate GPUs could be used – one for real-time OCT B-scan processing and one for deep learning inference of the layer boundaries. Peer-to-Peer (P2P) data transfer between separate GPU units is available and fast using NVLink interconnect, which supports transfer bandwidths as high as 300 GB/s, up to a 70% superior to PCIe transfer bandwidth.

Simultaneous B-scan and *en face* rendering of the OCT volume in real time allows the operator to: (1) ensure prior to acquisition that the position of the pupil is at an optimal plane to reduce pupil vignetting and B-scan cropping; (2) avoid saving scans containing large motion artifacts and blinks, thus improving the acquisition success rate; and (3) perform layer-specific optimization of defocus in large numerical-aperture instruments using adaptive optics [46–48]. Another direction that could be used to frame the work is that OCT is on the verge of transitioning from purely diagnostic applications to being used for image-guided surgeries. The capabilities shown here could provide real-time feedback to eye surgeons during common procedures such as retinal laser photocoagulation, corneal grafting, or anti-angiogenic injection treatment.

This method also poses great potential in OCTA, which is typically visualized *en face* and is very sensitive to motion and shadow artifacts [49]. Because adaptive optics was incorporated in this instrument, the *en face* images produced in real-time could also be utilized to compensate aberrations such as low-order defocusing, astigmatism, and coma prior to acquisition. By extending the method with a parallel software performing vessel segmentation in real time, we could produce instantaneous quantification of the vessel density or the avascular area in diseases like DR, yielding real-time visualization and quantification of imaging biomarkers of disease. In addition, real-time visualization of *en face* OCTA in tandem with sparse widefield scanning could assist the operator in navigating to regions of interest (e.g. locations of peripheral vitreous or choroidal neovascularization), and adapt the eccentricity, field of view and sampling density to visualize them with higher definition.

In summary, to the best of our knowledge, we have demonstrated real-time retinal layer segmentation by incorporating a highly optimized UNet in the OCT retinal scanner for the first time in literature. Embedding artificial intelligence in ophthalmic imaging system could potentially enable next generations of smart imagers that are capable of active image acquisitions.

Funding

Michael Smith Foundation for Health Research; Natural Sciences and Engineering Research Council of Canada; Canadian Institutes of Health Research; Research to Prevent Blindness; National Eye Institute (P30 EY010572) (R01 EY031331).

Disclosures

Acner Camino: Optovue, Inc (P). Marinko V. Sarunic: Seymour Vision (I). Yifan Jian: Seymour Vision (I). These potential conflicts of interest have been reviewed and managed by OHSU as well as Simon Fraser University. Other authors declare that there are no conflicts of interest related to this article.

References

1. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *J. Am. Med. Assoc.* **316**(22), 2402–2410 (2016).
2. J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.* **24**(9), 1342–1350 (2018).
3. Q. You, Y. Guo, J. Wang, X. Wei, A. Camino, P. Zang, C. J. Flaxel, S. T. Bailey, D. Huang, T. S. Hwang, and Y. Jia, "Detection of Clinically Unsuspected Retinal Neovascularization with Wide-Field Optical Coherence Tomography Angiography," *Invest. Ophthalmol. Visual Sci.* **60**, 3278 (2019).
4. J. Wang, T. T. Hormel, L. Gao, P. Zang, Y. Guo, X. Wang, S. T. Bailey, and Y. Jia, "Automated diagnosis and segmentation of choroidal neovascularization in OCT angiography using deep learning," *Biomed. Opt. Express* **11**(2), 927–944 (2020).
5. D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong, "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes," *J. Am. Med. Assoc.* **318**(22), 2211–2223 (2017).
6. P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks," *JAMA Ophthalmol.* **135**(11), 1170–1176 (2017).
7. R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," *Ophthalmology* **124**(7), 962–969 (2017).
8. M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning," *Invest. Ophthalmol. Visual Sci.* **57**(13), 5200–5206 (2016).
9. A. M. Hagag, A. D. Pechauer, L. Liu, J. Wang, M. Zhang, Y. Jia, and D. Huang, "OCT Angiography Changes in the 3 Parafoveal Retinal Plexuses in Response to Hyperoxia," *Ophthalmol. Retina* **2**(4), 329–336 (2018).
10. S. Taylor, J. M. Brown, K. Gupta, J. P. Campbell, S. Ostmo, R. V. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, S. J. Kim, J. Kalpathy-Cramer, and M. F. Chiang, Imaging and Informatics in Retinopathy of Prematurity Consortium, "Monitoring Disease Progression With a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning," *JAMA Ophthalmol.* **137**(9), 1022–1028 (2019).
11. A. Camino, M. Zhang, L. Liu, J. Wang, Y. Jia, and D. Huang, "Enhanced Quantification of Retinal Perfusion by Improved Discrimination of Blood Flow From Bulk Motion Signal in OCTA," *Trans. Vis. Sci. Tech.* **7**(6), 20 (2018).
12. C. S. Lee, A. J. Tying, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomed. Opt. Express* **8**(7), 3440–3448 (2017).
13. J. Xue, A. Camino, S. T. Bailey, X. Liu, D. Li, and Y. Jia, "Automatic quantification of choroidal neovascularization lesion area on OCT angiography based on density cell-like P systems with active membranes," *Biomed. Opt. Express* **9**(7), 3208–3219 (2018).
14. A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Express* **8**(8), 3627–3642 (2017).
15. P. Prentašić, M. Heisler, Z. Mammo, S. Lee, A. Merkur, E. Navajas, M. F. Beg, M. Šarunic, and S. Lončarić, "Segmentation of the foveal microvasculature using deep learning networks," *J. Biomed. Opt.* **21**(7), 075008 (2016).
16. D. Lu, M. Heisler, S. Lee, G. W. Ding, E. Navajas, M. V. Sarunic, and M. F. Beg, "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network," *Med. Image Anal.* **54**, 100–110 (2019).
17. A. Camino, Z. Wang, J. Wang, M. E. Pennesi, P. Yang, D. Huang, D. Li, and Y. Jia, "Deep learning for the segmentation of preserved photoreceptors on en face optical coherence tomography in two inherited retinal diseases," *Biomed. Opt. Express* **9**(7), 3092–3105 (2018).
18. R. Zhao, A. Camino, J. Wang, A. M. Hagag, Y. Lu, S. T. Bailey, C. J. Flaxel, T. S. Hwang, D. Huang, D. Li, and Y. Jia, "Automated drusen detection in dry age-related macular degeneration by multiple-depth, en face optical coherence tomography," *Biomed. Opt. Express* **8**(11), 5049–5064 (2017).
19. Z. Wang, A. Camino, M. Zhang, J. Wang, T. S. Hwang, D. J. Wilson, D. Huang, D. Li, and Y. Jia, "Automated detection of photoreceptor disruption in mild diabetic retinopathy on volumetric optical coherence tomography," *Biomed. Opt. Express* **8**(12), 5384–5398 (2017).

20. Z. Wang, A. Camino, A. M. Hagag, J. Wang, R. G. Weleber, P. Yang, M. E. Pennesi, D. Huang, D. Li, and Y. Jia, "Automated detection of preserved photoreceptor on optical coherence tomography in choroideremia based on machine learning," *J. Biophotonics* **11**(5), e201700313 (2018).
21. M. J. Heiferman and A. A. Fawzi, "Progression of subclinical choroidal neovascularization in age-related macular degeneration," *PLoS One* **14**(6), e0217805 (2019).
22. B. Wang, A. Camino, S. Pi, Y. Guo, J. Wang, D. Huang, T. S. Hwang, and Y. Jia, "Three-dimensional structural and angiographic evaluation of foveal ischemia in diabetic retinopathy: method and validation," *Biomed. Opt. Express* **10**(7), 3522–3532 (2019).
23. J. S. Schuman, M. R. Hee, C. A. Puliafito, C. Wong, T. Pedut-Kloizman, C. P. Lin, E. Hertzmark, J. A. Izatt, E. A. Swanson, and J. G. Fujimoto, "Quantification of Nerve Fiber Layer Thickness in Normal and Glaucomatous Eyes Using Optical Coherence Tomography: A Pilot Study," *Arch. Ophthalmol.* **113**(5), 586–596 (1995).
24. J. Loo, L. Fang, D. Cunefare, G. J. Jaffe, and S. Farsiu, "Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2," *Biomed. Opt. Express* **9**(6), 2681–2698 (2018).
25. L. de Sisternes, G. Jonna, M. A. Greven, Q. Chen, T. Leng, and D. L. Rubin, "Individual Drusen Segmentation and Repeatability and Reproducibility of Their Automated Quantification in Optical Coherence Tomography Images," *Trans. Vis. Sci. Tech.* **6**(1), 12 (2017).
26. Z. Mammo, M. Heisler, C. Balaratnasingam, S. Lee, D.-Y. Yu, P. Mackenzie, S. Schendel, A. Merkur, A. Kirker, D. Albani, E. Navajas, M. F. Beg, W. Morgan, and M. V. Sarunic, "Quantitative Optical Coherence Tomography Angiography of Radial Peripapillary Capillaries in Glaucoma, Glaucoma Suspect, and Normal Eyes," *Am. J. Ophthalmol.* **170**, 41–49 (2016).
27. A. Mishra, A. Wong, K. Bizheva, and D. A. Clausi, "Intra-retinal layer segmentation in optical coherence tomography images," *Opt. Express* **17**(26), 23719–23728 (2009).
28. R. Zawadzki, A. Fuller, D. Wiley, B. Hamann, S. Choi, and J. Werner, "Adaptation of a support vector machine algorithm for segmentation and visualization of retinal structures in volumetric optical coherence tomography data sets," *J. Biomed. Opt.* **12**(4), 041206 (2007).
29. P. A. Dufour, L. Ceklic, H. Abdillahi, S. Schroder, S. D. Dzanet, U. Wolf-Schnurbusch, and J. Kowal, "Graph-Based Multi-Surface Segmentation of OCT Data Using Trained Hard and Soft Constraints," *IEEE Trans. Med. Imaging* **32**(3), 531–543 (2013).
30. L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomed. Opt. Express* **8**(5), 2732–2744 (2017).
31. E. M. A. Anas, P. Mousavi, and P. Abolmaesumi, "A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy," *Med. Image Anal.* **48**, 107–116 (2018).
32. C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, "SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound," *IEEE Trans. Med. Imaging* **36**(11), 2204–2215 (2017).
33. T. Wang, J. Xiong, X. Xu, M. Jiang, Y. Shi, H. Yuan, M. Huang, and J. Zhuang, "MSU-Net: Multiscale Statistical U-Net for Real-time 3D Cardiac MRI Video Segmentation," in *MICCAI*, 2019.
34. S. S. M. Salehi, S. R. Hashemi, C. Velasco-Annis, A. Oualam, J. A. Estroff, D. Erdogmus, S. K. Warfield, and A. Gholipour, "Real-time automatic fetal brain extraction in fetal MRI by deep learning," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, (2018), 720–724.
35. Y. Jian, K. Wong, and M. Sarunic, "Graphics processing unit accelerated optical coherence tomography processing at megahertz axial scan rate and high resolution video rate volumetric rendering," *J. Biomed. Opt.* **18**, 026002 (2013).
36. J. Xu, K. Wong, Y. Jian, and M. Sarunic, "Real-time acquisition and display of flow contrast using speckle variance optical coherence tomography in a graphics processing unit," *J. Biomed. Opt.* **19**(2), 026001 (2014).
37. M. Cua, S. Lee, D. Miao, M. J. Ju, P. Mackenzie, Y. Jian, and M. Sarunic, "Retinal optical coherence tomography at 1 μ m with dynamic focus control and axial motion tracking," *J. Biomed. Opt.* **21**(2), 026007 (2016).
38. T. Zhang, A. M. Kho, and V. J. Srinivasan, "Improving visible light OCT of the human retina with rapid spectral shaping and axial tracking," *Biomed. Opt. Express* **10**(6), 2918–2931 (2019).
39. P. Mécé, V. Mazlin, J. Scholler, P. Xiao, J. Sahel, K. Grieve, M. Fink, and C. Boccara, "Real-time axial retinal motion tracking and correction for consistent high-resolution retinal imaging with Full-Field Time-Domain Optical Coherence Tomography (FFOCT)," *Invest Ophthalmol Vis Sci* **60**, 022 (2019).
40. W. Janpongsri, J. Huang, N. Ringo, W. D. J., M. V. Sarunic, and Y. Jian, "Pseudo-real-time retinal layer segmentation for high-resolution adaptive optics optical coherence tomography," arXiv 2004.05264 (2020).
41. A. Camino, R. Ng, J. Huang, Y. Guo, S. Ni, Y. Jia, D. Huang, and Y. Jian, "Depth-resolved optimization of a real-time sensorless adaptive optics optical coherence tomography," *Opt. Lett.* **45**(9), 2612–2615 (2020).
42. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, (Springer International Publishing, 2015), 234–241.
43. P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage* **31**(3), 1116–1128 (2006).

44. S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation," *Opt. Express* **18**(18), 19413–19428 (2010).
45. B. Keller, M. Draelos, G. Tang, S. Farsiu, A. N. Kuo, K. Hauser, and J. A. Izatt, "Real-time corneal segmentation and 3D needle tracking in intrasurgical OCT," *Biomed. Opt. Express* **9**(6), 2716–2732 (2018).
46. M. J. Ju, M. Heisler, D. Wahl, Y. Jian, and M. Sarunic, "Multiscale sensorless adaptive optics OCT angiography system for in vivo human retinal imaging," *J. Biomed. Opt.* **22**(12), 121703 (2017).
47. H. R. G. W. Verstraete, M. Heisler, M. J. Ju, D. Wahl, L. Bliet, J. Kalkman, S. Bonora, Y. Jian, M. Verhaegen, and M. V. Sarunic, "Wavefront sensorless adaptive optics OCT with the DONE algorithm for in vivo human retinal imaging [Invited]," *Biomed. Opt. Express* **8**(4), 2261–2275 (2017).
48. K. S. K. Wong, Y. Jian, M. Cua, S. Bonora, R. J. Zawadzki, and M. V. Sarunic, "In vivo imaging of human photoreceptor mosaic with wavefront sensorless adaptive optics optical coherence tomography," *Biomed. Opt. Express* **6**(2), 580–590 (2015).
49. A. Camino, Y. Jia, J. Yu, J. Wang, L. Liu, and D. Huang, "Automated detection of shadow artifacts in optical coherence tomography angiography," *Biomed. Opt. Express* **10**(3), 1514–1531 (2019).